„Rendering Endangered Lexicons Interoperable through Standards Harmonization": the RELISH Project

Helen Aristar-Dry
Eastern Michigan University, Ypsilanti, US
2000 Huron River Drive, Suite 104 Ypsilanti, MI 48197 United States
hdry@linguistlist.org

Sebastian Drude, Menzo Windhouwer
Max Planck Institute for Psycholinguistics
Wundtlaan 1, 6525 XD, Nijmegen, the Netherlands
{Menzo.Windhouwer, Sebastian.Drude}@mpi.nl

Jost Gippert, Irina Nevskaya
University of Frankfurt
Postfach 11 19 32 D-60054 Frankfurt, Germany
{gippert, nevskaya}@em.uni-frankfurt

Abstract

The RELISH project promotes language-oriented research by addressing a two-pronged problem: (1) the lack of harmonization between digital standards for lexical information in Europe and America, and (2) the lack of interoperability among existing lexicons of endangered languages, in particular those created with the Shoebox/Toolbox lexicon building software. The cooperation partners in the RELISH project are the University of Frankfurt (FRA), the Max Planck Institute for Psycholinguistics (MPI Nijmegen), and Eastern Michigan University, the host of the Linguist List (ILIT). The project aims at harmonizing key European and American digital standards whose divergence has hitherto impeded international collaboration on language technology for resource creation and analysis, as well as web services for archive access. Focusing on several lexicons of endangered languages, the project will establish a unified way of referencing lexicon structure and linguistic concepts, and develop a procedure for migrating these heterogeneous lexicons to a standards-compliant format. Once developed, the procedure will be generalizable to the large store of lexical resources involved in the LEGO and DoBeS projects.

Keywords: endangered lexica, standards harmonization, interoperability

1. Objectives of the project and its significance

When a lexicon constitutes the only record of a dying or already extinct language, it can contribute unique linguistic and cultural information to our store of scientific knowledge. Making it interoperable with other lexical data becomes a critical research priority. However, despite the support accorded to initiatives to develop digital standards for language documentation within both the US and Germany, there still exist major barriers to lexicon interoperability. The most significant barrier is that standards-setting bodies have arrived at different standards for format and markup on the two sides of the Atlantic. Additionally, within each national community, divergences exist in lexicon format and markup, in part because field linguists have hitherto relied on software which does not offer the linguist adequate support in choosing structural or linguistic categories.

The RELISH project started in 2009 and has been going for over two years by now. RELISH promotes language-oriented research by addressing a two-pronged problem: (1) the lack of harmonization between digital standards for lexical information in Europe and America, and (2) the lack of interoperability among existing lexicons of endangered languages, in particular those created with the Shoebox/Toolbox lexicon building software incorporating the Multi-Dictionary Formatter (MDF) which is a *de-facto* standard widely used in the linguistic community. The cooperation partners in the RELISH project are the University of Frankfurt (FRA), the Max Planck Institute for Psycholinguistics (MPI), and Eastern Michigan University, the host of the Linguist List (ILIT). The project aims at harmonizing key European and American digital standards whose divergence has hitherto impeded international collaboration on language technology for resource creation and analysis, as well as web services for archive access. Focusing on several lexicons of endangered languages, the project establishes a unified way of referencing lexicon structure and linguistic concepts, and develops a procedure for migrating these heterogeneous lexicons to a standards-compliant format. Once developed, the procedure will be generalizable to the large store of lexical resources involved in the LEGO (Lexicon Enhancement via the GOLD Ontology) and DoBeS (*Dokumentation bedrohter Sprachen* 'Documentation of Endangered Languages') projects. By now, the LEGO archive has 17 lexicons of endangered languages and over 3000 wordlists. The DoBeS lexica are stored in The Language Archive of the MPI Nijmegen which offers web services for over 200 endangered lexica, mostly in the LEXUS format.

The project is of significance both to the user community and to the organizations which support their research. Language data are central to a large scientific community, including anthropologists, archaeologists, historians, geneticists, sociologists, and linguists. Ensuring the interoperability of any individual lexicon exponentially increases its potential scientific contribution. The current harmonization of standards will streamline the future development of software tools and web services deployed in lexical research. Accordingly, the outcomes of the project add value to other projects already funded with public funds in Europe (e.g., LIRICS, CLARIN) and the US (E-MELD, the Data-Driven Ontology Project, GOLDComm, LEGO); and it contributes to the ongoing

effort of developers and funding agencies to make the most efficient use of scarce resources.

As a collaboration between two of the organizations that have been instrumental in promoting both endangered languages documentation and standards development in Europe and the US, this project provides impetus for other standards harmonization efforts, as well as offers the scientific research community flexible and integrated access to important new digital materials. Already now we see a cumulative effect of the harmonizing efforts of the project: The developers of WebLicht, a linguistic web services chaining tool, at the University of Tübingen are interested in a pivot format for lexica. There have been interested responses of leading researchers of endangered languages in Germany, England, US and Russia. Moreover, there is an increasing interest in the RELISH discussions about a proper exchange format in the ISO TC37 and CLARIN initiatives since yet there is no agreed format and it is widely agreed that ISO needs to recommend one.

2. Workflow

During the first two years of the project, ILIT and its consultants and MPI have pursued the top-down approach, harmonizing terminology and structure between LMF/DCR (ISO 24613:2008) and LIFT/GOLD, while FRA has led the work on the bottom-up approach, analyzing the attributes and values in the selected endangered languages lexicons. Top-down and bottom-up approaches go hand in hand, in continuous interaction by all project members.

The overall work is divided into four parts:

I. Harmonization of Terminology between GOLD and the ISO Data Category Registry.
II. Harmonization of Structure between LIFT/GOLD and LMF core + extensions.
III. Conversion of endangered language lexicons into the interchange format that results from standards harmonization.
IV. Roundtripping of lexicons between LEXUS and LEGO, as proof of interoperability.

For all the parts, the lexica are processed to fulfill the following tasks:

- Selection of lexical resources taking into account cross-linguistic research questions, richness and variability of resource structure and content, and accessibility,
- Conversion of selected resources into an XML format to achieve syntactic homogeneity,
- Import of resulting XML-based resources into LEXUS and the LEGO uploader.

As test lexicons we have chosen Wichita (Caddoan), Tuvan (Turkic), Chalkan (Turkic), Udi (North Caucasian) and Batsbi (North Caucasian) on the LEXUS side and Mocovi (Mataco-Guaicuru), Western Sisaala (Niger-Congo) and Fulfulde (Niger-Congo) on the LEGO side. On the one hand, these languages represent pairs of relatively closely related languages (Tuvan and Chalkan, Udi and Batsbi, Western Sisaala and Fulfilde) which enables us to test various historical and comparative search options; on the other hand, the test lexica belong to various language types and are spoken on different continents which gives us an opportunity to use them for verifying various cognitive and typological hypotheses and perspectives. Moreover, some of the test languages are spoken in the contiguous areas, or in close contacts with other languages of the same affiliation, which provides for searches for lexical loans or other traces of contacts between these languages.

2.1 Harmonization of Structure between LIFT/GOLD and LMF core + extensions

In choosing a harmonization schema for the lexicons, we considered both LIFT and TEI (then under consideration by CLARIN). As CLARIN had not selected an official format by January, 2010, the LIFT format was chosen by RELISH as a starting point. LIFT is the interchange format among the suite of lexicon tools created by SIL International, e.g., Toolbox, FLEX, WeSay. However, further harmonization efforts with CLARIN and related initiatives are likely to be necessary in the future.

The RELISH participants agreed on working out a RELISH interchange format based on LEGO's "LL-LIFT" format, a restricted version of the more general LIFT standard which still validates against LIFT. It was also agreed that the RELISH interchange format should be LMF-compliant taking into account the need for an agreed-on format in Europe where the LMF is gradually expanding in this function. LMF is a meta-specification without a standard XML serialization, whereas LIFT is defined by an XML schema. Investigation determined that a mapping of LIFT to LMF was relatively easy to achieve; and the resulting XML schema may be expanded to become a candidate for the official LMF serialization. This RELISH interchange format (see *Figure 1*) was defined to include ISOcat (the ISO DCR: www.isocat.org) mappings in the metadata (see section 2.2).

A routine for converting the existing LEGO and LEXUS lexica into the RELISH interchange format was created ensuring structural interoperability among the lexicons. Automatic export into the RELISH format will be added to the LEXUS dictionary creation tool and also added to the LEGO site. The LEGO site, still under development, can be seen at http://lego.linguistlist.org/.

2.2 Harmonization of Terminology between GOLD and the ISO Data Category Registry

By mapping on the ISOcat data categories, a harmonization of the used semantic categories is achieved. A suitable transformation of the concepts of the GOLD ontology into data categories was created: the GOLD XML version was converted to the Data Category Interchange Format (DCIF), the XML format needed for the upload to ISOcat by means of an XSL Transformation. On the other hand, an MDF data category selection including all the lexicon categories used in the Multi-Dictionary Formatter was created in the ISOcat data category registry. Both data category selections (GOLD and MDF) have been made public and are already widely used by the linguistic community for their resources.

Also after all MDF categories were put in the ISOcat registry as a separate set and made public in the first year of work, the following issues had to be dealt with:

1) the problem of the language assignment which is integrated into many MDF markers; most of them are thus in fact 'complex' categories;
2) the problem of dealing with the category of "status" of languages; most MDF categories are presented through sets of markers distinguished by the status of a language in which the respective information is encoded: vernacular (the documented linguistic variety), national (the state or official language of the country where this variety is spoken), regional (the language of broader communication in the area where the documented variety is spoken) and English (the language of scientific description and glossing, normally English, but also German, Russian, etc.);
3) thus, the category of status is a sociolinguistic category, and the ISOcat category registry does not have such a semantic domain.

```
<lmf:LexicalResource xmlns:lmf="http://www.lexicalmarkupframework.org/"
xmlns:dcr="http://www.isocat.org/ns/dcr" xmlns:tei="http://www.tei-c.org/ns/1.0">
    <lmf:GlobalInformation>
        <tei:f name="languageCoding" dcr:datcat="http://www.isocat.org/datcat/DC-2482">
            <tei:symbol value="ISO639-3"/>
        </></>
    <lmf:Lexicon>
        <lmf:LexicalEntry xml:id="le1">
            <tei:f name="originalID">
                <string>3</></>
            <lmf:Lemma type="Form">
                <lmf:FormRepresentation>
                    <tei:f name="text">
                        <tei:string xml:lang="fuh-Latn">aadaade</>
                    </></></></>
            <lmf:Sense>
                <lmf:Definition>
                    <lmf:TextRepresentation>
                        <tei:f name="text">
                            <tei:string xml:lang="eng-Latn"
                             >to promise, to convenant, to agree, to enter a contract</>
                        </></></></>
                <lmf:SenseExample>
                    <lmf:TextRepresentation>
                        <tei:f name="text">
                            <tei:string xml:lang="fuh-Latn"
                              >Laamdo aaɗeke adunaaru fuu heɓan barke saabe lenyoi Ibrahiima.</>
                        </>
                    </lmf:TextRepresentation>
                    <lmf:TextRepresentation>
                        <tei:f name="text">
                            <tei:string xml:lang="eng-Latn"
                              >God promised that the whole world would be blessed because of
                               Abraham's lineage.</>
                        </></></></>
                <tei:f name="partOfSpeech">
                    <tei:string>Verb</></>
                <tei:f name="grammaticalInfo">
                    <tei:string>Infinitive</></>
                <tei:f name="dialects">
                    <tei:vColl>
                        <tei:string>Jelgoore</>
                        <tei:string>Yaagaare</>
                        <tei:string>Gurmaare</>
                        <tei:string>Mooslire</></></></>
                <tei:f name="semanticCategory">
                    <tei:string>Theological Terms</></></></>
            <lmf:SenseRelation targets="le74 le1757" targetType="cf"/>
            <lmf:SenseRelation targets="le1755" targetType="synonym"/>
        </></></></>
```

Figure 1. Fulfulde instance snippet of the RELISH interchange format. To limit space end tags have been replaced by </>

Our approaches to these problems have mutated in the course of the project. First we have encoded the complex MDF categories as individual (complex) ISO data categories the way they are presented in the MDF standard; an indication to the simple categories they include has been made. However, it was later decided to introduce the category of "status" into the ISOcat category registry, and it became possible to deal with these categories as combinations of simple ones, which was implemented in the LEXUS import/export processes; this procedure was first applied to the Udi lexicon. A strategy of mapping to multiple categories was worked out in Nijmegen by introducing the "status" and "language" tags that can be mapped independently. As a consequence of this new procedure, we revised our MDF category selection in the ISOcat registry and introduced the missing simple categories which had been previously encoded as inherent parts of complex ones.

A chart with interrelations between the MDF, GOLD and "standard" ISOcat data categories was created; types of relations between the categories were established to be implemented in the Relation Registry and in mapping the categories in the process of lexicon import into the LEXUS and into the RELISH interchange formats.

definition (vernacular)	http://www.isocat.org/datcat/DC-3695	subClassOf	definition	http://www.isocat.org/datcat/DC-1972
definition (national)	http://www.isocat.org/datcat/DC-3709	subClassOf	definition	http://www.isocat.org/datcat/DC-1972
definition (regional)	http://www.isocat.org/datcat/DC-3710	subClassOf	definition	http://www.isocat.org/datcat/DC-1972
national language	http://www.isocat.org/datcat/DC-3702	partOf	definition (national)	http://www.isocat.org/datcat/DC-3709
regional language	http://www.isocat.org/datcat/DC-3703	partOf	definition (regional)	http://www.isocat.org/datcat/DC-3710
vernacular language	http://www.isocat.org/datcat/DC-3706	partOf	definition (vernacular)	http://www.isocat.org/datcat/DC-3695

Table 1. Triplets with semantic relations between ISOcat data categories

The Relation Registry also called RELcat being developed by the MPI, see (Schuurman & Windhouwer, 2011), will allow specification of (individual) relationships between data categories from the ISOcat DCR and possibly other concept registries. The chart describing the relationships between the MDF data categories, the GOLD data categories and other ISOcat data categories was imported into RELcat, which will allow tools to use these relationships in broadening and generalizing semantic searches. A small fragment of this chart is presented by *Table I* showing relations between the complex categories "definition (national)", "definition (regional)", and "definition (vernacular)" and their correlative simple categories "definition", "regional language", "national language", etc. An attempt to visualize these relations was made (see *Figure 2*).

Beside the development of RELcat, the RELISH project was also the prime motivator for more stable import/export facilities of the new LEXUS back-end. The LEXUS back-end has been rewritten to be based on an XML database management system. In this back-end new import and export facilities have been realized. For example, the import facility for MDF lexica has vastly improved in stability. And this facility now also includes data category references for standard MDF markers.

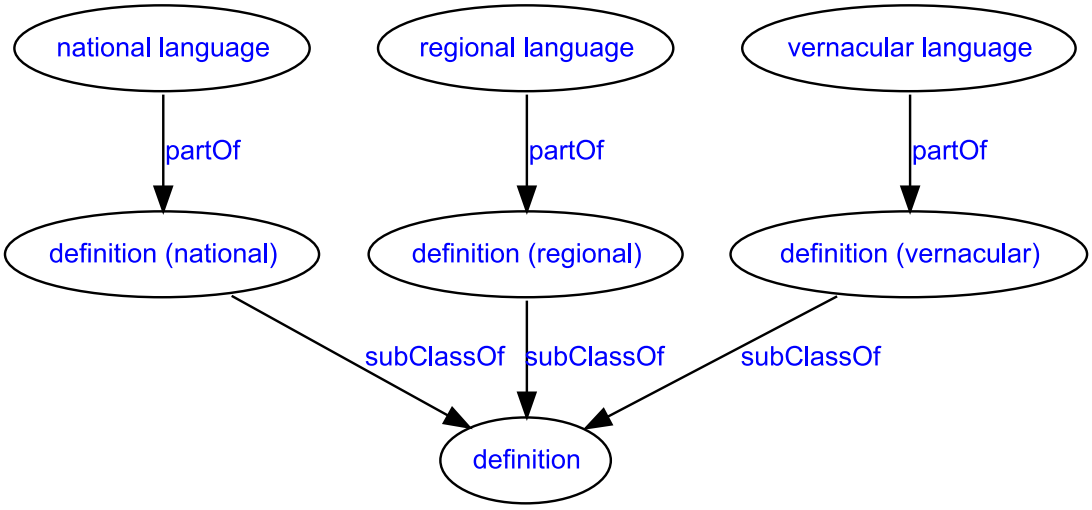

Figure 2. A fragment of the RELcat

3. Further to do

In the remaining time of the project we will concentrate on the following main tasks:

1) synchronization and update of controversial and complementary terminology with ISO/DCR community,
2) full serialization, full roundabout trip of all RELISH lexica: both sides should create an import routine and an export routine using the RELISH schema as an intermediate format.
3) creation of a demonstrator website where linguists can explore lexical resources, now interoperable, for cross-linguistic study; make all the tools available for the linguistic community.

References

GOLD: General Ontology for Linguistic Description: http://linguistics-ontology.org/.

ISO 24613. *Language resource management — Lexical markup framework (LMF).* (2008). Geneva: International Organization for Standardization.

LEGO: Lexicon Enhancement via the GOLD Ontology: http://linguistlist.org/projects/lego1.cfm

LEXUS: http://www.lat-mpi.eu/tools/lexus/

LIFT: Martin Hosken and Stephen McConnel (SIL Non-Roman Script Initiative (NRSI) and Language Software Development) *Lexicon Interchange Format. A Description* (version 0.15): http://lift-standard.googlecode.com/svn/trunk/

MDF: Coward, D. F. & Grimes, Ch. E. (2000). *Making Dictionaries: A guide to lexicography and the Multi-Dictionary Formatter*. Waxhaw, North Carolina: SIL International: http://www.sil.org/computing/shoebox/MDF_2000.pdf

Schuurman, I. & Windhouwer, M. A.. Explicit Semantics for Enriched Documents. What Do ISOcat, RELcat and SCHEMAcat Have To Offer? In: *2nd Supporting Digital Humanities conference (SDH 2011), 17-18 November 2011, Copenhagen, Denmark***.** Copenhagen, Denmark.